

# Before It Is Sent

*Toward prediction infrastructure for communication: forecasting outcomes before they happen, and keeping a record that can be checked*

EJDER DEMIR

*Finalit Research*

## Abstract

Weather, credit and elections all have forecasting infrastructure: predictions issued in advance, scored against what happens and improved by the scoring. Communication, the means by which most consequential outcomes between people and institutions are reached, has none. We ask whether a general language model can forecast the outcome of something before it is put out, and what evidence would show that it can. On 78 real, thread-opening emails with known outcomes (base rate of a reply 38.5%), all six systems we evaluate score above chance, five of them with intervals that exclude it. Our prediction-mode forecaster reaches an area under the ROC curve of 0.708 (95% interval 0.59 to 0.82). A single-sentence prompt to the same frontier model reaches 0.700 (0.58 to 0.82); the paired difference is +0.009 (-0.074 to +0.090). Messages ranked in the top quarter were answered 13 times in 19, against 3 in 19 for the bottom quarter. The signal, in other words, is already in the model, but it is a signal of order rather than of probability: read as probabilities, the raw scores are no more accurate than the base rate. The five-step scale our system reports is ordinally reliable: observed reply rates rise monotonically from 0% for messages called very unlikely to 63% for those called very likely. We argue that the scarce resource in this field is not forecasting capacity but outcome data linked to forecasts that were demonstrably made beforehand. We describe deployed infrastructure, which has so far issued 778 structured forecasts for 427 people, that produces exactly that: a per-message outcome question, a policy that counts silence as an answer, and a cryptographic seal that fixes every forecast at the moment it is made. We state the limits of this first study plainly, and they are substantial, and we set out a research agenda in which such a record becomes shared, verifiable infrastructure for individuals, organisations and institutions.

*Keywords:* prediction infrastructure, forecasting, language models, communication outcomes, calibration, AUC, prediction records, timestamping

---

## 1. Introduction

Most outcomes that matter between people and institutions are reached through communication: an application that is or is not shortlisted, a funding request that is or is not taken up, a price change that keeps or loses customers, a public consultation that does or does not change a policy. In each case the outcome is known only after the communication has gone out, when it can no

longer be changed. Other domains of consequence learned long ago to forecast before acting and to score forecasts afterwards: weather forecasting has been verified against outcomes for more than seventy years (Brier, 1950), and credit, insurance and elections all rest on forecasts that are issued in advance and checked. Communication has no equivalent. There is no forecast before, no score after, and no record that would let either improve.

This report is a first step toward that infrastructure. We treat the outcome as the object of interest and the communication as evidence about it. The question is empirical: given only what is known before a message is sent, can a model forecast whether it will achieve its purpose, and how well? A forecast that is never checked is an opinion; the study of human forecasting has shown that skill can only be established from a scored track record (Tetlock and Gardner, 2015). A forecast checked against what happened, at scale, is a measurement. The second half of this report is about how to make that checking trustworthy.

We make four contributions.

1. **A blind backtest protocol** for outcome forecasting on real messages, in which the model sees only the message and information available before sending, and outcomes are attached afterwards (Section 4).
2. **A comparison of six forecasting systems**, including a deliberately minimal one-sentence baseline, with bootstrap intervals and a paired test. The minimal baseline is not beaten on this set, and we report that as a finding rather than an inconvenience (Section 5).
3. **Evidence that a coarse ordinal scale is reliable where raw probabilities would mislead.** Reply rates rise monotonically across the five steps, but the absolute rates are not those a naive reading of the words would suggest. This motivates reporting steps rather than percentages until outcomes have calibrated them (Section 5.2).
4. **A design for a sealed prediction record**, now in production, which fixes each forecast at the moment it is made and, once its daily commitments are published, will let anyone verify that a forecast predates its outcome (Section 6), and an agenda for growing it into shared prediction infrastructure (Section 7).

We separate three things throughout. *What the evidence shows*: a preliminary association, on one sender's email, between model rankings and observed replies. *What the infrastructure establishes*: a way to commit to forecasts before outcomes and to evaluate them afterwards. *What we are pursuing*: forecasting that generalises across people, kinds of communication and outcomes. Only the first is a result of this report.

## 2. Related work

*Predicting responses to messages.* Whether a message is answered has been studied directly. Kooti et al. (2015) analysed reply behaviour at the scale of a large webmail service and showed how load and conversation history shape whether and when people reply. Yang et al. (2017) characterised reply behaviour in enterprise email and trained models to predict it from features of the message, its sender and its recipients. Althoff et al. (2014) showed that the success of a written request for help can be predicted from its text, and identified the linguistic factors involved. These

studies learn from large labelled corpora drawn from one platform and one kind of message. We ask a different question: whether a general model, with no task-specific training, can forecast an outcome before a message is sent, for any kind of message, and how such forecasts can be checked.

*Forecasting with language models.* Halawi et al. (2024) build a retrieval-augmented system that approaches the accuracy of human crowds on questions from forecasting platforms. Those questions concern public events. Our setting is private and local: the event is the consequence of one person's message, and no crowd forecast exists to compare against.

*Simulating people.* A line of work uses language models to stand in for human respondents, from survey populations (Argyle et al., 2023) to interacting generative agents (Park et al., 2023). Simulation produces a distribution of synthetic responses. We do not simulate. We issue a single forecast about a real event and wait for the world to resolve it, which makes our claims checkable in a way that simulated populations are not.

*Models as judges.* Language models are widely used to grade text (Zheng et al., 2023). Grading agreement with human raters is a different target from outcome prediction: a well-written message can still go unanswered. We evaluate only against outcomes.

*Calibration and scoring.* Modern neural networks are often miscalibrated (Guo et al., 2017), although large language models carry useful information about their own uncertainty (Kadavath et al., 2022). Proper scoring rules (Brier, 1950; Gneiting and Raftery, 2007) reward honest probabilities, and their decomposition separates reliability from resolution (Murphy, 1973). With few resolved outcomes per category, we report discrimination (AUC; Hanley and McNeil, 1982) and ordinal reliability, and defer probabilistic scoring until the record supports it.

*Commitment and timestamping.* Forecast evaluation is vulnerable to hindsight: predictions revised, selected or reconstructed after the fact. Preregistration addresses this for scientific claims (Nosek et al., 2018); leakage between training and evaluation addresses it for models (Kaufman et al., 2012). Haber and Stornetta (1991) showed how to timestamp documents so that their existence at a given time can be proven; hash trees (Merkle, 1987) let many commitments be published as one. Our record applies these ideas to individual forecasts.

### 3. Problem

A sender is about to send a message  $x$  with pre-send context  $c$  (who it is for, what it is meant to achieve). There is an outcome event  $E$ , such as *the recipient replies*, which will later be observed as true or false. A forecaster maps  $(x, c)$  to an ordered judgment of how likely  $E$  is. Two properties matter. **Discrimination**: messages for which  $E$  occurred should receive higher forecasts than those for which it did not. **Reliability**: among messages given the same forecast, the frequency of  $E$  should match what the forecast claims. Discrimination can be established with modest samples. Reliability, in the strong sense of matching stated probabilities, needs many outcomes per forecast level, which is precisely what the record in Section 6 is built to accumulate.

Two features make the problem harder than it first appears. First, the relevant event differs by message: a reply for an introduction, a shortlisting for an application, a renewal for a customer notice. In production the event is therefore derived per message rather than taken from a fixed list.

Second, silence is an outcome. A system that records only the answers people volunteer will over-represent successes and report itself as more accurate than it is.

## 4. Data and protocol

### 4.1 Messages and outcomes

The evaluation set consists of 78 emails drawn from a single sender's mailbox export. Only messages that opened a thread were kept, because a reply to a reply proves little; automated recipients, and very short or very long messages, were excluded. The outcome is whether anyone other than the sender later wrote in the same thread. 30 messages received a reply and 48 did not, of which 5 received a brief refusal, counted here as a failure. The base rate is therefore 38.5%. Messages were used exactly as sent; context fields contain only what was known before sending.

### 4.2 Blind judging

Each system saw the message and its pre-send context and nothing else. Outcomes were held in a separate file and joined to the scores only after all judgments were complete. Because judging is blind, a later correction to a label changes the evaluation without requiring any message to be judged again.

### 4.3 Systems

**Plain prompt.** A single instruction to the model: *"You will be shown a message someone sent. Estimate the probability, from 0 to 100, that it achieved what it was for (for example, that the recipient replied). Answer with a single integer and nothing else."* This is the minimal baseline: no method, no structure, the same underlying model.

**Readiness judge.** Finalit's earlier production judge, which infers the type of message, its recipient and its purpose, lists what is holding it back, and returns a readiness score from 0 to 100 and a verdict of ready or not ready. We evaluate it on three models.

**Models.** All systems run on frontier language models from a single provider, used through its public interface with no fine-tuning. We refer to them as Model A and Model C, two large models from successive generations, and Model B, a mid-sized model from the same generation as Model A. The plain prompt, prediction mode and its variant all use Model C, so differences between them reflect method, not model. Exact model identifiers and versions are available to reviewers on request.

**Prediction mode.** The current production forecaster. It names the outcome event for the message, places it on a five-step scale from very unlikely to very likely, and returns a score from 0 to 100. A variant adds a three-line frame (what the outcome turns on, what the forecast assumes, and the earliest observable sign) to test whether added explanation changes the forecast.

### 4.4 Statistics

We report the area under the ROC curve (Hanley and McNeil, 1982; Fawcett, 2006), which is the probability that a randomly chosen replied message is scored above a randomly chosen unreplied one, with ties counting one half. Intervals are 95% percentile bootstrap intervals over 2,000 res-

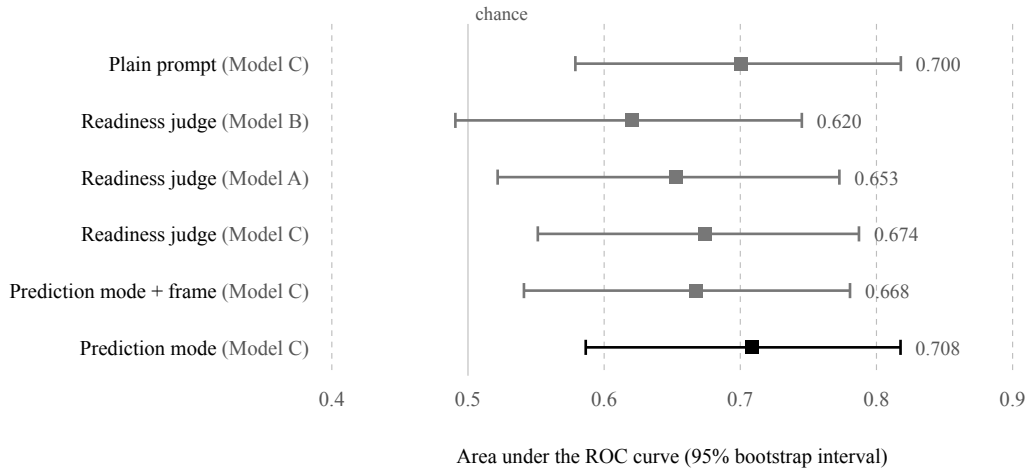
amples of messages (Efron and Tibshirani, 1993). The comparison between prediction mode and the plain prompt uses a paired bootstrap of 4,000 resamples, so both systems are scored on identical resamples; this is a resampling alternative to the test of DeLong et al. (1988) for correlated curves. Significance against chance uses a label-permutation test: outcome labels are shuffled 10,000 times while scores are held fixed, under the null hypothesis that scores carry no information about outcomes. Scores were fixed before any outcome was attached, because judging was blind. The full computation is short enough to audit line by line and is reproduced from the raw scores in the accompanying scripts.

## 5. Results

### 5.1 Every system carries signal; the minimal one is not beaten

**Table 1.** Discrimination on 78 messages (30 replies). Intervals are 95% bootstrap intervals.

| <i>System</i>           | <i>Model</i> | <i>AUC</i> | <i>95% interval</i> |
|-------------------------|--------------|------------|---------------------|
| Plain prompt            | Model C      | 0.700      | [0.58, 0.82]        |
| Readiness judge         | Model B      | 0.620      | [0.49, 0.75]        |
| Readiness judge         | Model A      | 0.653      | [0.52, 0.77]        |
| Readiness judge         | Model C      | 0.674      | [0.55, 0.79]        |
| Prediction mode + frame | Model C      | 0.668      | [0.54, 0.78]        |
| Prediction mode         | Model C      | 0.708      | [0.59, 0.82]        |

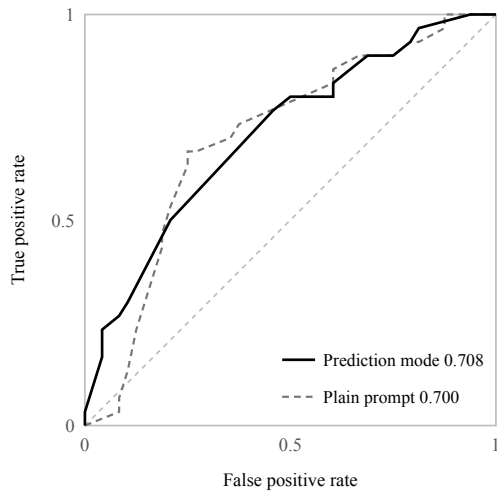


**Fig. 1.** Area under the ROC curve for each system, with 95% bootstrap intervals. Every lower bound except one is above chance. The intervals overlap heavily: on 78 messages, differences of a few hundredths cannot be resolved.

All six systems score above chance, and for five of them the interval excludes it (Table 1, Fig. 1). Prediction mode scores highest at 0.708, and the plain prompt scores 0.700. The paired difference is +0.009, with an interval from -0.074 to +0.090: no detectable difference. The two systems also

agree closely on the order in which they rank the messages (Spearman 0.80; Fig. 2). Against chance, both results are significant (label-permutation test, 10,000 permutations:  $p = 0.001$  for prediction mode and 0.002 for the plain prompt). In practical terms, of the 19 messages prediction mode ranked highest, 13 were answered (68%); of the 19 it ranked lowest, 3 were (16%). This is a statement about prediction, not intervention: it shows that the ranking carries information about outcomes, not that acting on a forecast changes the chance of a reply.

We read this as the central empirical result of the study. A frontier model, asked one plain question, already holds a real and non-trivial signal about whether a message will be answered. Prompt engineering moved discrimination within the noise. Whatever advantage a forecasting system can build over time will not come from its instructions, which any competitor can approximate in an afternoon. It will have to come from something the model does not have.



**Fig. 2.** ROC curves for prediction mode and the plain prompt on the same 78 messages. The curves cross repeatedly; neither dominates.

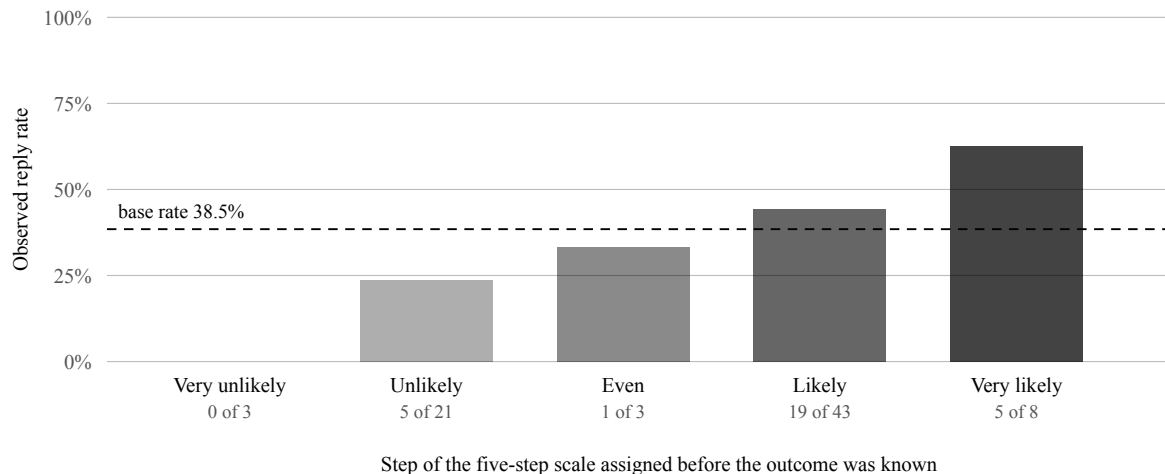
**Table 2.** Outcomes by forecast step (prediction mode).

| <i>Step</i>   | <i>Messages</i> | <i>Replied</i> | <i>Rate</i> |
|---------------|-----------------|----------------|-------------|
| Very unlikely | 3               | 0              | 0%          |
| Unlikely      | 21              | 5              | 24%         |
| Even          | 3               | 1              | 33%         |
| Likely        | 43              | 19             | 44%         |
| Very likely   | 8               | 5              | 63%         |

Base rate 38.5%

### 5.2 The five-step scale is ordinally reliable

Prediction mode reports a step, not a number. Table 2 and Fig. 3 show what happened to the messages at each step. Reply rates rise monotonically: none of the 3 messages called very unlikely received a reply, 5 of 21 called unlikely did, 19 of 43 called likely did, and 5 of 8 called very likely did. Two of the five steps hold only three messages each, so the ordering at those steps is suggestive rather than established.



**Fig. 3.** Observed reply rate for messages at each step, with counts. The order is right at every step. The levels are not what the words alone suggest: a message called likely was answered 44% of the time, above the base rate of 38.5% but well short of what most readers would take "likely" to mean.

This pattern is the reason the product does not print percentages. The ordering is reliable; the absolute levels are not yet calibrated, and on these counts could not be. A percentage would claim a precision the evidence does not support. We prefer a coarse claim that holds to a precise one that does not, and we treat calibration of the levels as something to be earned from resolved outcomes, per context, as the record grows.

### 5.3 Ranking without calibration

Both systems also produce a score from 0 to 100, which can be read as a probability. Read that way, neither is useful. The Brier score (Brier, 1950) of the plain prompt is 0.229 and that of prediction mode 0.256, against 0.237 for a forecaster that always states the base rate. The corresponding skill scores are +0.03 and -0.08: no better than climatology (Murphy, 1973). The models order messages well and price them badly. Their mean stated probabilities (50% and 55%) sit well above the observed rate of 38.5%.

This is, in our view, the most practically important result in the report. A system that printed these numbers as percentages would present a forecast with no probabilistic skill as if it were precise. The evidence supports a claim about order, which an ordinal scale expresses honestly, and does not yet support a claim about probability. Probabilities have to be earned from resolved outcomes, context by context, which is the purpose of the record described in Section 6.

### 5.4 Equal AUC does not mean equal forecasts

The readiness judge illustrates why aggregate discrimination is an insufficient release criterion. On Model A it called 3 of 78 messages ready; on Model B, 27; on Model C, 30. Models B and C reach similar AUCs (0.620 and 0.674), yet agree on the ready or not-ready verdict for only 57 of 78 messages (73%). Two systems can rank a population equally well while giving a particular person opposite answers. For a product whose output is a verdict on one message, per-message agreement matters as much as AUC, and we now report both whenever a model or prompt changes.

### 5.5 Explanations should not move the forecast

Adding the three-line frame changed the step assigned to 11 of 78 messages; 67 were identical and 77 were within one step. Discrimination moved from 0.708 to 0.668, within the noise. We treat this as a regression test that a production change must pass: an explanation added for the user's benefit should not silently change what is being forecast.

### 5.6 Repeatability

To estimate run-to-run variation, one message was judged ten times at each of three reasoning-effort settings (Model A). The readiness score varied with a standard deviation of 6.1 points at the default setting, 5.4 at low and 3.2 at medium, and the verdict was identical in all 30 runs. This is a probe, not a study: a single message cannot establish repeatability across messages, and a larger repeated-measures design is planned.

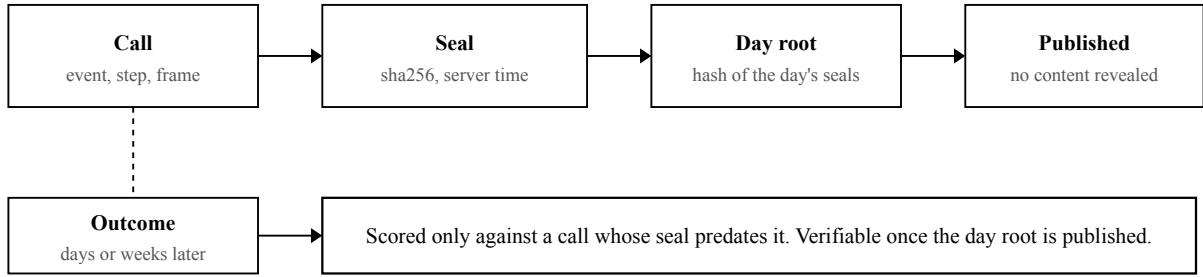
## 6. From forecasts to a record

If the signal is in the model, the question becomes what a forecasting system can accumulate that a model cannot. Our answer is a record: forecasts made before outcomes, joined to what then happened, at a scale and breadth that no published dataset covers. Such a record is only as valuable as it is trustworthy. Three design decisions follow.

**The question is part of the forecast.** With every judgment, the forecaster also writes the outcome question for that message: what to ask, the possible answers in order, and how long to wait before asking is fair. A fixed taxonomy of message types would fail on the first unfamiliar one; deriving the question from the message is the same act as deriving the event being forecast.

**Silence is an answer.** Every outcome question offers a way to say that nothing happened, at the same weight as a positive answer, and answers such as *still waiting* postpone the question rather than resolve it. Without this, the record would contain mainly the successes people chose to report and would overstate accuracy.

**Every forecast is sealed.** Since 11 October 2026, each forecast is fixed at the moment it is logged (Fig. 4). The server, not the client, assigns the time. A SHA-256 digest is computed over the forecast's identifier, owner, time, verdict, claim and structured prediction. After sealing, those fields cannot be changed through the application; later information, such as what the sender did next, remains writable without disturbing the seal. A seal held only by the operator is not, by itself, independent evidence: an operator with administrative access to the database could in principle alter a record and recompute its digest. Independent verifiability therefore requires a further step, which is planned. Each day's seals are combined into a single root, following Haber and Stornetta (1991), and that root must be published somewhere the operator does not control. Once a root is published, any later change to a forecast from that day becomes detectable by anyone, without the forecast itself being revealed. Sealing is in production; publication of roots is not yet, and until it is, the record should be read as sealed but not independently verifiable. Forecasts logged before 11 October 2026 were sealed on that date and can be evidenced only from then.



**Fig. 4.** The sealed record. A forecast is sealed when it is made; each day's seals are combined into a published root; outcomes arrive later and are scored only against forecasts whose seal predates them. Independent verification depends on publishing the roots outside the operator's control, which is planned.

**Deployment.** By 11 October 2026 the forecaster had issued 778 structured forecasts for 427 people, across the full range of things they were about to send. The steps were distributed as follows: very unlikely 4.8%, unlikely 43.1%, even 22.1%, likely 29.3% and very likely 0.8%. The forecaster rarely makes its most confident call: 6 times in 778. In-product outcome collection began in October 2026; resolved outcomes are not yet numerous enough to analyse, and we do not report them.

The product is designed to show people their own record under each forecast, but only once at least three of their own forecasts of that kind have resolved, and only as counts (for example, four checked and three as called). No figure is shown that the person's own outcomes do not support.

## 7. Toward prediction infrastructure

The study above uses one kind of communication and one outcome. The problem it belongs to is far broader, and we set out here what infrastructure for it would require. Parts of this are in production; the rest is a research agenda, and we mark the difference.

**A common unit.** Infrastructure begins with a standard object. Ours is the sealed forecast: an outcome event named for the specific item, a step on an ordinal scale, the time after which the outcome can fairly be known, and a seal. The unit is the same whether the item is an email, an application, an announcement or a policy proposal. *In production.*

**Any outcome.** The event is derived per item rather than drawn from a fixed taxonomy, so the same infrastructure covers a reply, a shortlisting, a grade, a renewal or a vote, and the outcome question is generated with the forecast. *In production; evaluated here only for replies.*

**Calibration by context.** Section 5.2 shows the ordering of the scale is already right and its levels are not yet known. As resolved outcomes accumulate, the levels can be calibrated separately for each context in which enough outcomes exist, and a forecast can carry its own track record. Individual records are being introduced in the product; calibration across many records is *future work*.

**Verifiability by third parties.** Published daily roots would allow anyone holding a forecast and its seal to show that it predates its outcome, and would allow accuracy claims to be audited rather than taken on trust (Haber and Stornetta, 1991; Merkle, 1988). An append-only log of pub-

lished roots, held outside the operator's control, is what turns a sealed record into an independently verifiable one. *Sealing in production; publication of roots is planned.*

**From individuals to institutions.** The same record serves at every scale. An individual learns how their own calls have resolved. A team or an organisation accumulates a shared record across everything it sends. An institution could forecast the outcome of a public communication before issuing it, and later show that the forecast was made in advance. Each scale adds outcomes, and each outcome improves calibration for everyone in that context. *Individuals in production; organisational and institutional records are future work.*

**Open questions.** Four questions define the agenda. Does discrimination hold for new people, whose messages played no part in developing the forecaster? Does it hold for other kinds of communication and other outcomes? Does it hold across models, so that the record rather than any one model carries the result? And do accumulated outcomes improve forecasts, in discrimination and in calibration, beyond what a model achieves without them?

## 8. Discussion

The comparison in Section 5.1 could be read as a weakness: our forecaster does not beat a sentence. We read it as a statement about where value lies. Forecasting capacity is becoming a commodity, available to anyone who can write a prompt. What is not available is a large body of forecasts that were demonstrably made before their outcomes, across the full variety of things people send, joined to what actually happened. That body of evidence is what would allow levels to be calibrated per context, systematic errors to be found and corrected, and claims of accuracy to be made at all. It cannot be scraped, because the forecasts did not exist before they were made, and once daily roots are published it cannot be backfilled without detection.

A record is infrastructure, not yet an advantage. The claim that matters, and that this report does not test, is that forecasts conditioned on accumulated outcomes outperform forecasts made without them, particularly in contexts where a general model has no comparable information. That is the question our next report is designed to answer: prospectively, on forecasts sealed before their outcomes, across many people, under an evaluation protocol declared in advance (Nosek et al., 2018).

The reliability result in Section 5.2 points the same way. The ordering of the scale is already right; its levels are not yet known. Each resolved outcome narrows that gap for the context it belongs to. The forecasts can become more precise only as the record grows, and we think precision should be reported only at the rate it is earned.

## 9. Limitations

The evidence in this report is limited, and we prefer to say so directly. (i) The evaluation set is small: 78 messages, 30 positive. Intervals are wide and differences of several hundredths are unresolvable. (ii) All messages come from one sender's mailbox, so the results may not transfer to other senders, styles or relationships. (iii) The only outcome measured is a reply to an email; most of what people send has other outcomes and other channels. (iv) The production prompts were developed with this set available, so the scores of the Finalit systems may be optimistic relative to

the plain baseline, which was not tuned. A held-out set is required before any claim of superiority could be made, and we make none. (v) We do not compare against human forecasters. (vi) Reasoning models are not deterministic; Section 5.6 measures this on only one message. (vii) The record described in Section 6 is new: it holds 778 forecasts, but its resolved outcomes are not yet enough to report discrimination or calibration on live use, and we report neither. (viii) The seal is currently held by the operator; until daily roots are published externally, the integrity of the record rests on the operator's controls rather than on independent verification.

## 10. Ethics and data

The evaluation messages are private correspondence and are not released. Recipients are not identified in this report, and only aggregate results are shown. In production, the record stores forecasts and outcome answers for the account that made them; the published daily roots reveal neither content nor identity.

## 11. Conclusion

A language model can forecast, better than chance, whether a message will achieve its purpose, and it can do so from a single plain question. It ranks well and prices badly, so the scale on which our system reports forecasts is ordinal, and it orders outcomes correctly. Frontier models supply the signal; it is available to anyone. What no model supplies is a record of forecasts sealed before their outcomes and checked against them afterwards. Built at the scale of individuals, organisations and institutions, that record becomes what communication has never had: prediction infrastructure. We have begun to build it.

## References

- Althoff, T., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2014). How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 12–21. doi:10.1609/icwsm.v8i1.14547
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351. doi:10.1017/pa.n.2023.2
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
- DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837–845. doi:10.2307/2531595
- Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. doi:10.1016/j.patrec.2005.10.010
- Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. doi:10.1198/016214506000001437

- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70, 1321–1330. arXiv:1706.04599
- Haber, S., and Stornetta, W. S. (1991). How to time-stamp a digital document. *Journal of Cryptology*, 3(2), 99–111. doi:10.1007/BF00196791
- Halawi, D., Zhang, F., Chen, Y.-H., and Steinhardt, J. (2024). Approaching human-level forecasting with language models. arXiv:2402.18563
- Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36. doi:10.1148/radiology.143.1.7063747
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., et al. (2022). Language models (mostly) know what they know. arXiv:2207.05221
- Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 1–21. doi:10.1145/2382577.2382579
- Kooti, F., Aiello, L. M., Grbovic, M., Lerman, K., and Mantrach, A. (2015). Evolution of conversations in the age of email overload. In *Proceedings of the 24th International Conference on World Wide Web*, 603–613. doi:10.1145/2736277.2741130
- Merkle, R. C. (1988). A digital signature based on a conventional encryption function. In *Advances in Cryptology: CRYPTO '87*, Lecture Notes in Computer Science 293, 369–378. doi:10.1007/3-540-48184-2\_32
- Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595–600. doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. doi:10.1073/pnas.1708274114
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. doi:10.1145/3586183.3606763
- Tetlock, P. E., and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. New York: Crown.
- Yang, L., Dumais, S. T., Bennett, P. N., and Awadallah, A. H. (2017). Characterizing and predicting enterprise email reply behavior. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 235–244. doi:10.1145/3077136.3080782
- Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*. arXiv:2306.05685